研究与开发

CR-NOMA中基于深度确定策略梯度的能效优化策略

张云
（南京交通职业技术学院，江苏 南京 211100）

摘　要：利用认知无线电非正交多址接入（cognitive radio non-orthogonal multiple access，CR-NOMA）技术可缓解频谱资源短缺问题，提升传感设备的吞吐量。传感设备的能效问题一直制约着传感设备的应用。为此，针对CR-NOMA中的传感设备，提出基于深度确定策略梯度的能效优化（deep deterministic policy gradient-based energy efficiency optimization，DPEE）算法。DPEE算法通过联合优化传感设备的传输功率和时隙分裂系数，提升传感设备的能效。将能效优化问题建模成马尔可夫决策过程，再利用深度确定策略梯度法求解。最后，通过仿真分析了电路功耗、时隙时长和主设备数对传感能效的影响。仿真结果表明，能效随传感设备电路功耗的增加而下降。此外，相比于基准算法，提出的DPEE算法提升了能效。

关键词：传感设备；能量采集；认知无线电非正交多址接入；能效；深度确定策略梯度
中图分类号：TN929.53
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024146

Deep deterministic policy gradient-based energy efficiency optimization algorithm for CR-NOMA

ZHANG Yun
Nanjing Communications Institute of Technology, Nanjing 211100, China

Abstract: Cognitive radio non-orthogonal multiple access (CR-NOMA) technology was used to alleviate the shortage of spectrum resource, and improve the throughput of sensor devices. But the energy efficiency problem had been restricting the application of sensor devices. Therefore, for CR-NOMA, deep deterministic policy gradient-based energy efficiency optimization (DPEE) algorithm was proposed. By jointly optimizing the transmission power and time slot splitting coefficient, the energy efficiency of sensor devices was improved. The energy efficiency optimization problem was modeled as a Markov decision process, and it was solved by the deep deterministic policy gradient (DDPG) method. Finally, the influence of circuit power consumption, time slot durations and number of main devices on energy efficiency were analyzed. The simulation results show that the energy efficiency decreases as the circuit power consumption of sensor device increases. In addition, compared with other algorithms, the proposed algorithm improves energy efficiency.

收稿日期：2024−01−20；**修回日期：**2024−03−05
基金项目：江苏省青蓝工程青年骨干教师项目
Foundation Item: Jiangsu Province Blue-Project Young Backbone Teachers Project

Key words: sensor device, energy harvesting, cognitive radio non-orthogonal multiple access, energy efficiency, deep deterministic policy gradient

0　引言

随着对监测物理环境需求的不断增长，传感设备已在多个领域广泛使用[1]。例如，在工业环境中，将传感设备安装在危险或人难以接近的地方，由传感设备作为感测点，采集工业环境数据。由于给传感设备更换电池成本较高，常从射频信号中采集能量，进而给传感设备补给能量[2]。因此，针对支持传感设备的无线网络，设计兼顾能量和频谱有效的通信协议是非常重要的。

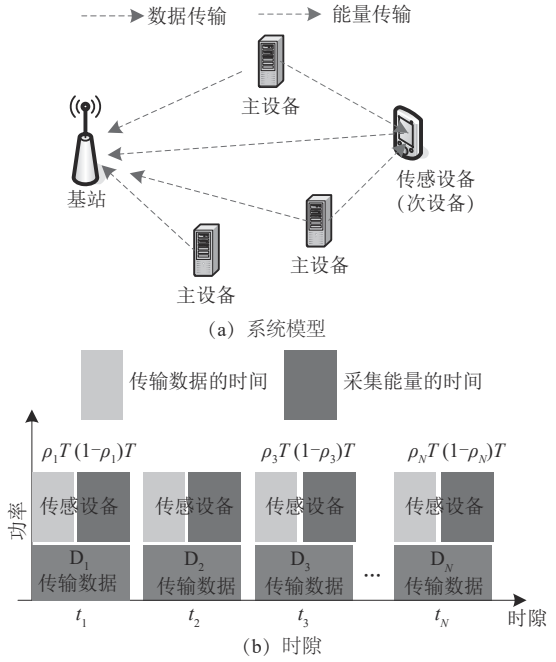
认知无线电非正交多址接入（cognitive radio non-orthogonal multiple access，CR-NOMA）[3]技术是提高频谱效率和吞吐量的有效技术。CR-NOMA 技术通过认知无线电策略，允许设备依据网络状态，调整采集能量和通信参数[4]。然而，在动态网络和能量环境下，动态地调整传感设备的传输功率是一项挑战性工作。

深度强化学习（deep reinforcement learning，DRL）能够从动态环境中学习，并进行智能决策，其对动态环境有较强的鲁棒性。文献[5]联合深度确定策略梯度（deep deterministic policy gradient，DDPG）和凸优化算法求解最优协议和功率分配策略，进而提升能量采集效率和最大化网络吞吐量。类似地，文献[6]针对 CR-NOMA 网络，提出基于 DDPG 的吞吐量优化（DDPG-throughput，DDPG-T）算法。然而，上述工作在优化吞吐量时，未考虑传感设备内部电路的电路功耗。实际上，内部电路的电路功耗对传感设备能效存在不可忽略的影响。

为此，针对 CR-NOMA 系统，提出基于深度确定策略梯度的能效优化（deep deterministic policy gradient-based energy efficiency optimization，DPEE）算法。该算法考虑了传感设备的电路功耗。传感设备的能效等于传感设备所获取的速率与所消耗的总能量之比[7]。主要工作如下：建立能效优化模型；利用 DDPG 算法求解问题；分析电路功耗、时隙以及主设备数对能效的影响。性能分析结果表明，提出的 DPEE 算法能够有效地提高设备的能效。

1　系统模型

非正交多址接入（non-orthogonal multiple access，NOMA）技术和认知无线电（cognitive radio，CR）技术都允许传感设备以频谱共享方式接入信道，提高了频谱利用率。基于 NOMA 和 CR 两种技术融合的 CR-NOMA 网络能进一步优化次网络的能效效率和传感设备的速率和，已在健康医疗、运输与物流等行业中应用。为此，考虑基于 CR-NOMA 的系统模型，如图 1（a）所示，系统由一个基站、基站覆盖的 N 个主设备（primary device，PD）和一个传感设备构成，传



（a）系统模型

（b）时隙

图 1　系统模型

感设备作为次设备，能以频谱共享方式接入信道。这些主设备构成PD集$U=\{\mathrm{D}_i|1\leqslant i\leqslant N\}$，其中$\mathrm{D}_i$表示第$i$个PD。它们采用时分多址接入技术，每个PD在其固定时隙内传输数据，且每个时隙的时长为T，总时隙数M，时隙如图1（b）所示。

给PD分配时隙的原则是将第k个时隙t_k分配给D_i：

$$i=\left((k-1)\oplus N\right)+1 \tag{1}$$

其中，$\oplus$表示异或操作，k表示时隙索引号，且$1\leqslant k\leqslant M$。

传感设备采用CR-NOMA，PD为主设备，传感设备为次设备。因此，传感设备能在PD的传输时隙内，占用频带资源向基站传输数据。

此外，传感设备具有采集能量的功能。可从PD向基站的上行射频信号中采集能量。为此，传感设备将每个时隙划分为两个部分，一部分用于传输数据，另一部分用于采集能量。对于第k个时隙t_k，传感设备传输数据的时间为$\rho_k T$；传感设备利用剩余的时间$(1-\rho_k)T$采集能量，其中$\rho_k\in[0,1]$为时隙分裂系数。

令g_e表示传感设备至基站的信道增益，主设备D_i至基站和传感设备的信道增益分别表示为g_i和$g_{i,e}$，如图1（a）所示。

令$E_{\max}$表示传感设备具有最大的能量，传感设备的能量变化示意图如图2所示。用E_k表示传感设备在时隙t_k所具有的能量。因此，传感设备在时隙t_k所消耗的能量应不高于E_k，即$\bar{E}_k=\rho_k T\left(P_k^e+Q\right)\leqslant E_k$,其中$\bar{E}_k$表示消耗的能量，$P_k^e$表示在时隙$t_k$时传感设备的传输功率，$Q$表示传感设备电路功耗。

传感设备在时隙t_k所采集的能量为：

$$\hat{E}_k=(1-\rho_k)T\eta\Omega_k^i\left|g_{i,e}\right|^2 \tag{2}$$

其中，η表示能量采集系数，Ω_k^i表示在时隙t_k传感设备从主设备D_i接收的信号功率。

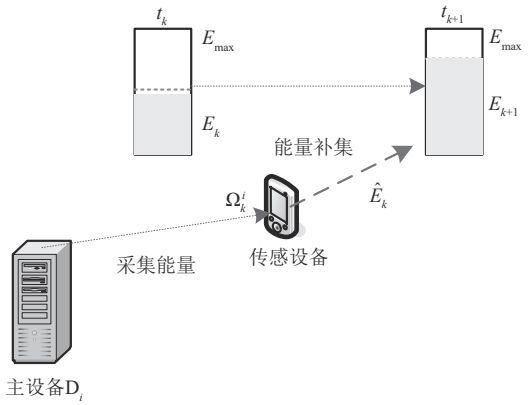


图2　传感设备的能量变化示意图

因此，传感设备在时隙t_{k+1}存有的能量为：

$$E_{k+1}=\min\left\{E_k+\hat{E}_k-\bar{E}_k,E_{\max}\right\} \tag{3}$$

2　问题描述

DPEE算法旨在最大化传感设备的能效。将能效定义为传感设备所传输的数据量与传感设备功耗之比。由于采用NOMA技术，基站端利用串行干扰消除（successive interference cancellation，SIC）技术[8]解码各设备信号。依据NOMA技术[9]的解码规则，优先解码信号最强的用户，并视其他用户信号为噪声。因此，可将传感设备的能效定义为：

$$\Gamma_{\mathrm{EE}}\left(\rho_k,P_k^e\right)=\frac{\rho_k\,\mathrm{lb}\left(1+\dfrac{P_k^e\left|g_e\right|^2}{1+\Omega_k^i\left|g_{i,e}\right|^2}\right)}{P_k^e+Q} \tag{4}$$

其中，$\mathrm{lb}\left(1+\dfrac{P_k^e\left|g_e\right|^2}{1+\Omega_k^i\left|g_{i,e}\right|^2}\right)$表示传感设备的速率。式（4）中分子表示所传输的数据量，分母表示所消耗的功率。

据此，最大化传感设备能效问题（以下简称优化问题）可表述为：

$$\mathrm{P0}:\ \max_{\rho_k,P_k^e}\ \Gamma_{\mathrm{EE}}\left(\rho_k,P_k^e\right) \tag{5}$$

$$\text{s.t.}\quad 0\leqslant P_k^e\leqslant P_{\max} \tag{5a}$$

$$0\leqslant \rho_k\leqslant 1 \tag{5b}$$

$$\rho_k T\left(P_k^e+Q\right)\leqslant E_k \tag{5c}$$

$$E_{k+1}=\min\left\{E_k+\hat{E}_k-\bar{E}_k,E_{\max}\right\} \tag{5d}$$

其中，$P_{\max}$表示允许传感设备的最大传输功率，约束项（5a）将传感设备的传输功率限定在特定范围内，约束项（5b）限定了时隙分裂系数。

考虑到优化问题的非凸性[10]，利用DDPG算法求解该优化问题。

3　DPEE算法

3.1　DDPG算法简述

DDPG算法的目的在于决策一个动作，使该动作产生最大的动作效益[11]：

$$a^*(s)=\arg\max_a\ Q(s,a) \tag{6}$$

其中，a表示动作，s表示状态，$Q(s,a)$表示在当前状态s下实施动作a所产生的效益函数。

与Q-学习算法不同，DDPG算法无须建立存储$Q(s,a)$的表，它是利用神经网络最大化$Q(s,a)$。此外，DDPG算法能处理连续动作空间。具体而言，DDPG算法包含以下4类神经网络[12]。

（1）Actor网络：该网络以状态s为输入，输出一个动作$\mu\left(s|\omega_\mu\right)$，其中$\omega_\mu$为Actor网络参数。

（2）目标Actor网络：该网络输出为$\mu_t\left(s|\omega_{\mu_t}\right)$，其中$\omega_{\mu_t}$为目标Actor网络参数。

（3）Critic网络：该网络以状态s和动作a为输入，输出为相应的效应值$Q\left(s,a|\omega_c\right)$，其中$\omega_c$为Critic网络参数。

（4）目标Critic网络：该网络输出为$Q_t\left(s,a|\omega_{c_t}\right)$，$\omega_{c_t}$为目标Critic网络参数。

DDPG算法的训练过程如下[13]。

（1）从用户角度，由于Actor网络提供决策，Actor网络是DDPG算法最重要的一个训练环节。为了增加样本的随机性，在Actor网络中添加噪声。因此，输出的动作为：

$$a(s)=\mu\left(s|\omega_\mu\right)+n \tag{7}$$

其中，n表示噪声。

（2）Actor网络更新。引入四元组$\langle s,a,r,s_\rangle$，其中r表示在状态s下实施动作a所产生的奖励，$s_$表示下一个状态。

通过最大化Q值训练Actor网络。利用Actor网络和Critic网络参数最大化目标函数：$J\left(\omega_\mu\right)=Q\left(s,a=\mu\left(s|\omega_\mu\right)|\omega_c\right)$。

再通过计算目标函数关于ω_μ的偏微分，利用梯度上升法更新Actor网络参数：

$$\nabla_{\omega_\mu}J\left(\omega_\mu\right)=\nabla_a Q\left(s,a|\omega_c\right)\nabla_{\omega_\mu}\mu\left(s|\omega_\mu\right) \tag{8}$$

（3）Critic网络的更新。通过最小化损失函数更新Critic网络：

$$L=\left(y-Q\left(s,a|\omega_c\right)\right)^2 \tag{9}$$

其中，y表示状态值的目标值：

$$y=r+\gamma Q_t\left(s\_,\mu_t\left(s\_|\omega_{\mu_t}\right)|\omega_{c_t}\right) \tag{10}$$

其中，γ为折扣率。

（4）目标网络更新。利用软更新方式更新Actor和Critic的目标网络：

$$\begin{cases}\omega_{c_t}\leftarrow\tau\omega_c+(1-\tau)\omega_{c_t}\\ \omega_{\mu_t}\leftarrow\tau\omega_\mu+(1-\tau)\omega_{\mu_t}\end{cases} \tag{11}$$

其中，τ表示软更新参数。

（5）回收经验池。DDPG算法将历史的元组信息存储于回收池。当需要更新网络时，就以随机方式从回收池提取一些元组数据，用于训练网络。

（6）学习率。Actor和Critic网络采用相同的学习率。不采用固定的学习率，而是采用动态衰减学习率：$\eta=\eta_{\text{init}}\times\delta^{\text{Episode}}$，其中$\eta_{\text{init}}$表示初始的学习率，$\delta$表示衰减因子。

3.2　基于DDPG求解问题

基于DDPG求解P0问题如图3所示。DDPG算法主要由状态、动作以及奖励函数构成。接下

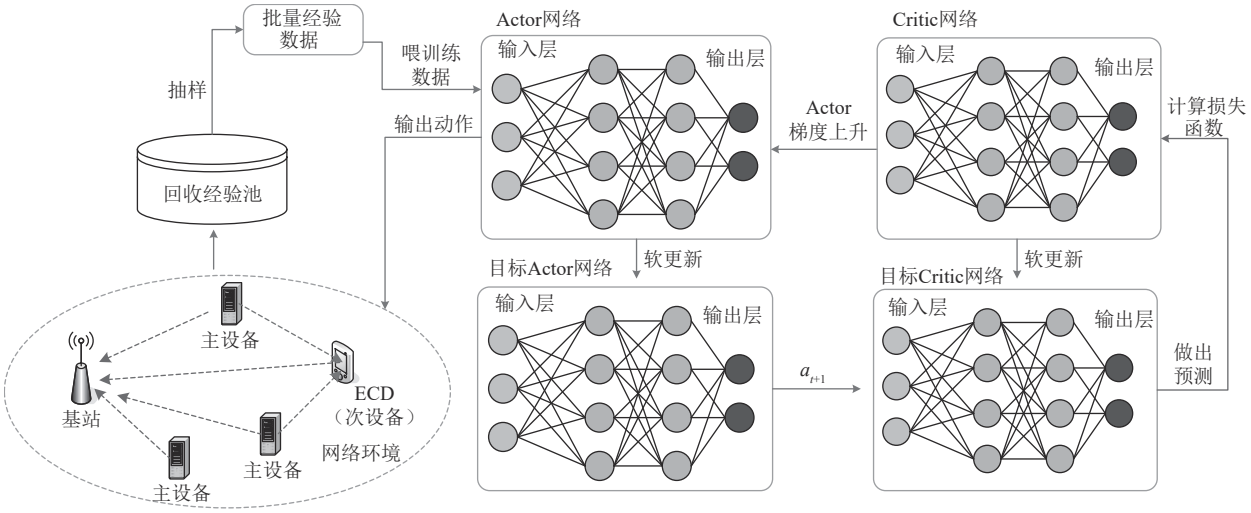


图3　基于DDPG求解P0问题

来，对状态、动作以及奖励函数进行设计。

（1）状态

将所有信道增益和传感设备的能量表述状态。因此，时隙t状态空间可表述为：

$$\bar{s}_k=\left\{E_k,\left|g_e\right|^2,\left|g_{i,e}\right|^2,\left|g_i\right|^2\right\} \tag{12}$$

（2）动作

回顾问题P0，通过优化时隙分裂系数ρ_k和传输功率P_k^e，最大化传感设备的能效。因此，将ρ_k和P_k^e作为动作，通过调整动作，最大化传感设备的能效。

$$a_t=\left\langle \rho_k,P_k^e\right\rangle \tag{13}$$

若传感设备不传输数据，只采集能量，$\rho_k=0$。反之，若只传输数据不采集能量，$\rho_k=1$。

（3）奖励函数

奖励函数是深度强化算法的重要部分。通过奖励引导动作朝最大化能效方向调整。因此，将传感设备的能效作为其奖励值，如式（14）所示：

$$r_t=\Gamma_{\mathrm{EE}}\left(\rho_k,P_k^e\right) \tag{14}$$

3.3　DPEE算法执行步骤

算法1给出DPEE算法的执行流程。首先，随机化初始Actor和Critic网络的权重；再对Actor和Critic的目标网络进行随机化；随后进入迭代环节。

算法1　DPEE算法

初始化：随机化初始Actor和Critic网络的权重

　　随机化Actor和Critic的目标网络

进入迭代环节

for Episode k=1 to K do

　为探索动作产生随机噪声

　根据环境观察初始状态s_k

　for 每个Episode中步数j=1 to N do

　　依据当前策略，选择一个动作a_k，并计算能效。

　　执行动作a_k，计算相应的奖励r_k，再进入下一个状态

　　将$\langle s_k,a_k,r_k,s_{k+1}\rangle$存储于经验回收池

　　从经验回收池中随机抽取mini-batch数据

　　依式（9）计算目标值

$$y_k=r_k+\gamma Q_k\left(s_{k+1},\mu_t\left(s_{k+1}|\omega_{\mu_t}\right)|\omega_{c_t}\right)$$

　　最小化损失函数，并更新Critic网络

$$L\left(\omega_c\right)=\sum_k\left|y_k-Q\left(s_{k+1},\mu_t\left(s_{k+1}|\omega_c\right)\right)\right|^2$$

　　利用随机梯度下降法，更新Actor网络

$$\nabla\omega_\mu J\left(\omega_\mu\right)=\nabla_a Q\left(s_{k+1},a_k|\omega_c\right)$$
$$\nabla_{\omega_\mu}\mu\left(s_{k+1}|\omega_\mu\right)$$

更新 Actor 和 Critic 目标网络：

$$\begin{cases}\omega_{c_t}\leftarrow\tau\omega_c+(1-\tau)\omega_{c_t}\\ \omega_{\mu_t}\leftarrow\tau\omega_\mu+(1-\tau)\omega_{\mu_t}\end{cases}$$

end for

end for

4　性能分析

4.1　仿真环境

为了更好地分析 DPEE 算法性能，利用 Python3.8 建立仿真平台。基站位于二维平面上的原点位置，N 个主设备和一个传感设备随机分布于基站的覆盖范围区域。采用 Rayleigh 衰落模型[14-15]。具体的仿真参数见表 1。

表 1　仿真参数

仿真参数	值
主设备数量	2~6
训练元组数	64
传感设备电池的最大容量/J	0.2
传感设备的最大传输功率/dBm	23
能量采集系数	0.9
时隙时长/s	1
折扣率	0.85

Actor 网络由 2 层组成，第 1、2 层的神经元分别为 256、128 个。Critic 网络由 2 层组成，第 1、2 层的神经元分别为 1 024、512 个。输入层采用 ReLU 激活函数，输出层采用 Tanh 激活函数。Actor 网络和 Critic 网络的学习率均为 0.005。总的 Episode 次数为 14 000，每次迭代中执行 100 步，mini-batch 的大小为 64。此外，为了更好地分析 DPEE 算法的性能，选择 2 个基准算法。

（1）DDPG-T[6]：DDPG-T 算法通过联合优化传感设备的传输功率和时隙，提升传感设备的吞吐量。DPEE 算法与 DDPG-T 算法不同之处：DPEE 算法通过优化传输功率和时隙，最大化传感设备的能效；而 DDPG-T 算法旨在最大化传感设备的吞吐量。

（2）Stochastic algorithm（STOA）[5]：STOA 算法通过联合优化传感设备的充电时间和传输功率，提升吞吐量。DPEE 算法与 STOA 算法不同之处：STOA 算法未考虑传感设备电路功耗，此外，STOA 算法旨在最大化吞吐量，而不是优化能耗。

4.2　折扣率和学习率对平均奖励的影响

折扣率 γ 对算法的平均奖励有重要的影响，其取值直接影响了平均奖励，其中 Episode 表示迭代次数，总共迭代了 14 000 次。折扣率对平均奖励的影响如图 4 所示，给出了 γ=0.6、0.85 和 0.99 共 3 种情况下的平均奖励。从图 4 可知，γ 取 0.85 时的平均奖励最高，且收敛性能也较好，迭代至约 10 次后，就收敛。而 γ=0.6 和 γ=0.99 时的平均奖励随 Episode 快速下降。为此，DPEE 算法将 γ 设置为 0.85。值得说明的是：只选择 0.6、0.85 和 0.99 进行仿真测试，没有穷举更多值去测试。

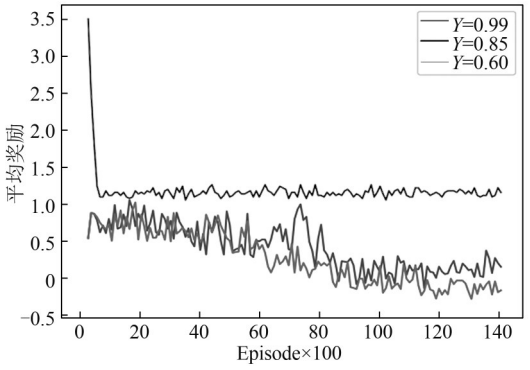


图 4　折扣率对平均奖励的影响

学习率对平均奖励的影响如图 5 所示，其中 η_{init}=0.01。衰减因子 δ=0.999、0.9 和 0.2。从图 5 可知，在 η_{init} 相同的条件下，δ=0.999、0.9 和 0.2 的 3 种情况下的平均奖励相差不大。从整体看，δ=0.9 下的平均奖励总体优于 δ=0.999 和 0.2 两种情况，但其平均奖励波动较大。而 δ=0.999 情况下平均奖励随迭代次数的波动较小，变化较平稳。

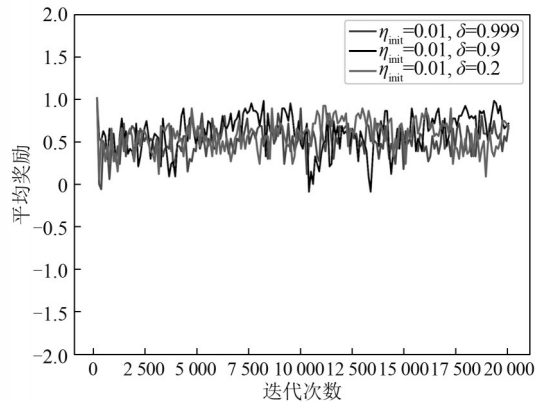


图5　学习率对平均奖励的影响

4.3　传感设备所采集的能量

本节分析DDPG-T、STOA和DPEE算法中传感设备所采集的能量（以下简称采-能）。

首先分析传感设备电路功耗（Q）和主设备数（N）对采-能的影响，采-能随主设备数的变化情况如图6所示。其中Q从0至0.03W变化，N等于2、6。从图6可知，Q值对采-能影响较小。在Q从0增加至0.03W过程中，采-能略有增加。原因在于：Q值越大，传感设备所消耗的能量越大（$\rho_k T\left(P_k^e+Q\right)$），传感设备就需要采集更多能量。由于$Q$值较小，$Q$值对采-能影响较小。

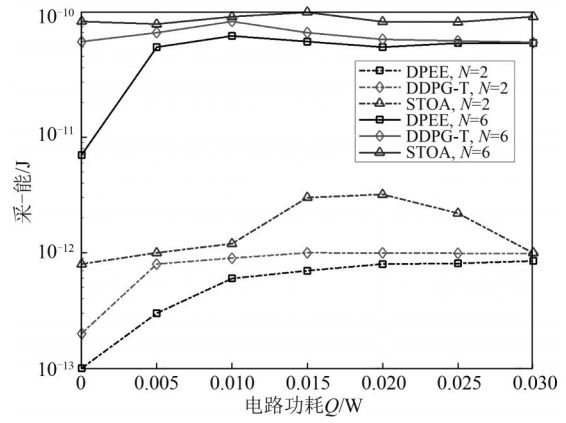


图6　采-能随主设备数的变化情况

相比之下，N值对采-能影响较大。以DPEE算法为例，在Q=0.03 W时，DPEE算法在N=2时的采-能约为10^{-12} J，而当N值增加至6时，采-能提升至5×10^{-11} J。原因在于：N值越大，传感设备采集能量的机会越多。

此外，相比于DDPG-T和STOA算法，DPEE算法中传感设备的采-能最少。这说明DPEE算法中传感设备的能耗最低。当设备有较充足能量时就无须采集能量。4.3节将分析传感设备的能效性能，将充分证明DPEE算法能效方面上的优势。STOA算法采集的能量最多，这说明STOA算法能耗速度高于DPEE算法和DDPG-T算法。当Q从0增至0.01时，STOA算法所采集的能量有较大幅度的增加。原因在于：当电路功耗较大时，传感设备能耗速度过快，电量低，需要采集更多能量来补给设备能量。但是当Q增至0.02后，若再继续提升电路功耗，设备所采集的能量就开始下降。高的电路功耗，需要补给更多能量，但这需要更长的充电时间。可能没有更长时间给设备充电，设备需要传输数据（传输数据需要时间），这导致所采集的能量下降。

采-能随时隙时长的变化情况如图7所示，其中N=2。从图7可知，T的增加有利于提升传感设备的采-能。原因在于：T值越大，主设备的时间窗口越长，传感设备可采集能量的时间越多。因此，能够采集更多能量。

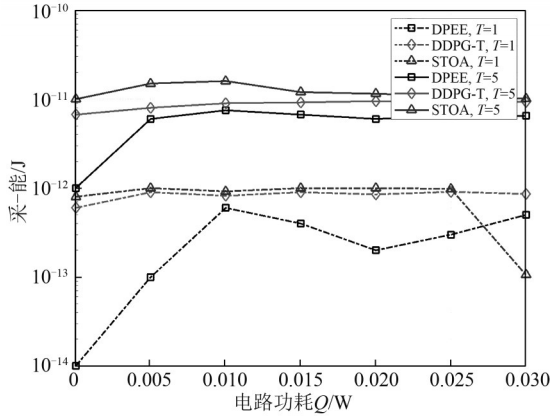


图7　采-能随时隙时长的变化情况

4.4　能效

本节分析传感设备的能效随主设备数的变化

情况，如图 8 所示，其中 Q 从 0 至 0.03 W 变化，N 等于 2、6。

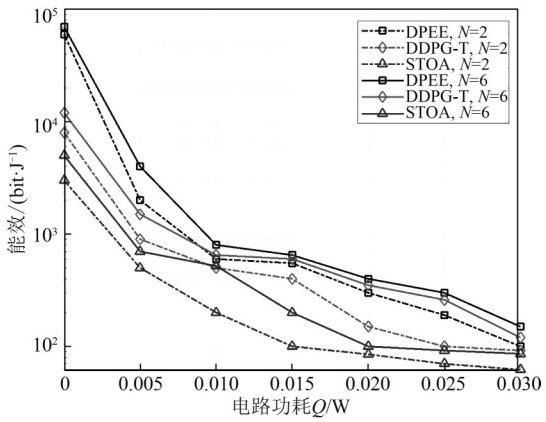


图 8　能效随主设备数的变化情况

从图 8 可知，能效随传感设备电路功耗的增加而下降。原因在于：Q 值越大，传感设备消耗的能量越多，依式（4）的能效定义，能效就下降。相比于 DDPG-T 和 STOA 算法，DPEE 算法提升了能效。一方面，DPEE 算法以提升能效为目的，通过优化传输功率和分配采集能量时隙的比例（ρ_k），提高能效。另一方面，DPEE 算法考虑了传感设备的电路功耗。图 5 也验证了传感设备电路功耗（Q）对能效存在不可忽略的影响。此外，增加主设备数（增大 N 值）可提升能效。原因在于：主设备数越多，传感设备获取采集能量的机会越多，有利于补给传感设备的能量。

4.5　吞吐量

DDPG-T、STOA 和 DPEE 算法中传感设备的吞吐量随时隙时长的变化情况如图 9 所示，其中 N=2，T=1, 8。

从图 9 可知，传感设备电路功耗（Q）值对吞吐量的影响较小。在 Q 从 0 增加至 0.03 W 过程中，吞吐量的波动较小。原因在于：Q 值增加或者减少，只会直接影响到传感设备能耗速度。此外，T 值的增加使吞吐量下降。这主要因为：增加 T 值，传感设备获取更长的采集能量的时间，

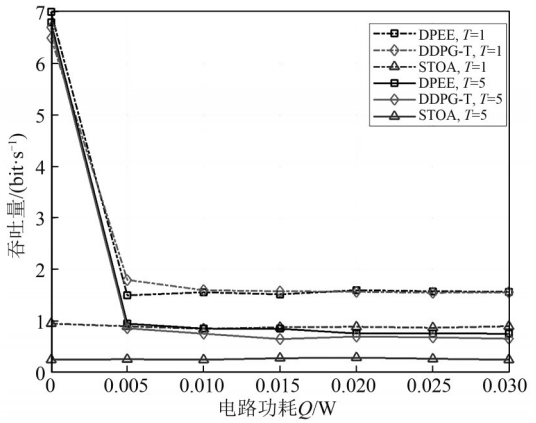


图 9　吞吐量随时隙时长的变化情况

缩短了传输数据时长。图 9 的数据也证实：T 值越大，所采集的能量越多。换而言之，采集的能量与吞吐量成反比。

DPEE、DDPG-T 和 STOA 算法得到图 8 结果所产生的运行时间见表 2。通过记录运行时间分析算法的复杂度。运行时间越长，算法越复杂。从表 2 可知，DDPG-T 算法的运行时间最长，达到 59.53 s；而本文提出的 DPEE 算法的运行时间介于 DDPG-T 算法与 STOA 算法之间。STOA 算法只考虑优化传感设备的能耗，没有考虑能效，算法相对简单，所以 STOA 算法的运行时间最短。

表 2　运行时间

算法	运行时间/s
DPEE	56.27
DDPG-T	59.53
STOA	52.43

5　结束语

在传感设备能采集能量的环境下，本文研究了传感设备的能效问题。通过优化传感设备的传输功率和时隙分裂系数，最大化传感设备的能效。先建立优化传感设备能效问题，再利用凸优化和 DDPG 法求解。最后，分析了电路功耗 Q 以

及时隙时长、主设备数对传感设备能效的影响。分析结果表明，传感设备能效随电路功耗Q的增加而下降，而时隙时长和主设备数对传感设备能效影响较小。此外，相比于DDPG和STOA算法，DPEE算法提升了传感设备的能效。

参考文献：

[1] 范绍帅, 吴剑波, 田辉. 面向能量受限工业物联网设备的联邦学习资源管理[J]. 通信学报, 2022, 43(8): 65-77.
FAN S S, WU J B, TIAN H. Federated learning resource management for energy-constrained industrial IoT devices[J]. Journal on Communications, 2022, 43(8): 65-77.

[2] GUO S T, SHI Y W, YANG Y Y, et al. Energy efficiency maximization in mobile wireless energy harvesting sensor networks[J]. IEEE Transactions on Mobile Computing, 2018, 17(7): 1524-1537.

[3] 乔宇航, 贺玉成, 杨键泉, 等. 两级中继选择的协作CR-NOMA网络性能分析[J]. 信号处理, 2020, 36(2): 196-202.
QIAO Y H, HE Y C, YANG J Q, et al. Performance analysis for cooperative CR-NOMA with two-stage relay selection[J]. Journal of Signal Processing, 2020, 36(2): 196-202.

[4] DING Z G, SCHOBER R, POOR H V. A new QoS-guarantee strategy for NOMA assisted semi-grant-free transmission[J]. IEEE Transactions on Communications, 2021, 69(11): 7489-7503.

[5] LI L Y, XU H B, MA J, et al. Joint EH time and transmit power optimization based on DDPG for EH communications[J]. IEEE Communications Letters, 2020, 24(9): 2043-2046.

[6] DING Z G, SCHOBER R, POOR H V. No-pain No-gain: DRL assisted optimization in energy-constrained CR-NOMA networks[J]. IEEE Transactions on Communications, 2021, 69(9): 5917-5932.

[7] WANG Z, GE L N, LI T S, et al. Energy efficiency maximization strategy for Sink node in SWIPT-enabled sensor-cloud based on optimal stopping rules[J]. China Communications, 2021, 18(1): 222-236.

[8] 颜晓娟, 安康, 张千锋, 等. 面向用户随机分布的NOMA-卫星通信网络性能分析[J]. 电讯技术, 2020, 60(2): 153-158.
YAN X J, AN K, ZHANG Q F, et al. Performance analysis of NOMA-based satellite networks with randomly deployed users[J]. Telecommunication Engineering, 2020, 60(2): 153-158.

[9] 褚翼飞, 杨新杰. 支持多个多播组业务的多中继协作NOMA系统中断性能研究[J]. 电信科学, 2023, 39(7): 35-45.
CHU Y F, YANG X J. Outage performance of multi-relay collaborative NOMA systems supporting multiple multicast services[J]. Telecommunications Science, 2023, 39(7): 35-45.

[10] BOYD S P, VANDENBERGHE L. Convex optimization[M]. Cambridge, UK: Cambridge, 2004.

[11] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge, Mass.: MIT Press, 1998.

[12] 陈清林, 邝祝芳. 基于DDPG的边缘计算任务卸载和服务缓存算法[J]. 计算机工程, 2021, 47(10): 26-33.
CHEN Q L, KUANG Z F. Task offloading and service caching algorithm based on DDPG in edge computing[J]. Computer Engineering, 2021, 47(10): 26-33.

[13] 司鹏搏, 吴兵, 杨睿哲, 等. 基于DDPG三维无人机路径规划[J]. 高技术通讯, 2022, 32(10): 1049-1057.
SI P B, WU B, YANG R Z, et al. A DDPG algorithm to UAV path planning in 3D[J]. Chinese High Technology Letters, 2022, 32(10): 1049-1057.

[14] 樊湖. 无线数据通信快速组网方法及组网成功率计算[J]. 无线电工程, 2021, 51(2): 104-110.
FAN H. Fast networking method and networking success probability calculation of wireless data communication system[J]. Radio Engineering, 2021, 51(2): 104-110.

[15] 刘德稳, 招继炳, 盛冬发, 等. 基于Rayleigh阻尼模型的层间隔震结构上下部选择[J]. 湖南科技大学学报（自然科学版）, 2022, 37(3): 27-34.
LIU D W, ZHAO J B, SHENG D F, et al. Research model selection of mid-story isolated structure based on Rayleigh damping[J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2022, 37(3): 27-34. .

[作者简介]

张云 （1989− ），女，南京交通职业技术学院讲师，主要研究方向为智能交通、控制算法。